

# 面向教育场景的视觉大模型优化与应用

许跃蓬<sup>1,2</sup>, 徐柴迪<sup>3</sup>, 郭晋军<sup>1</sup>, 姜云桥<sup>4</sup>, 王仕嘉<sup>1</sup>, 刘垚<sup>1,2,5</sup>

1. 华东师范大学数据科学与工程学院, 上海 200062;
2. 先进计算与智能工程国家级重点实验室, 江苏 无锡 214000;
3. 上海孔棣科技有限公司, 上海 201204;
4. 摩尔线程智能科技(成都)有限公司, 四川 成都 610031;
5. 软硬件协同设计技术与应用教育部工程研究中心(华东师范大学), 上海 200062

## 摘要

随着人工智能技术的迅速发展, 大语言模型在许多领域取得了显著成功, 但在教育领域的应用仍面临多模态数据处理困难、回答准确性不足以及信息载体单一等问题。为解决这些问题, 提出了一种视觉教育大模型 VELM。VELM 基于多模态公开教育数据集及专用教育数据集进行训练, 并结合模型优化技术, 不仅提升了模型在教育场景中的回复质量, 而且实现了计算资源的优化与降低。同时, 使用了 RAG 技术确保生成内容的准确性和丰富度。在部署与应用方面, VELM 通过 Dify 平台实现了灵活的多端部署, 支持微信小程序、Web 云端平台以及本地化部署 3 种形式, 满足了不同教育场景中的多样化需求。评测实验表明, VELM 在标准评测数据集(如 Mathvista、OCRBench 和 MMMU 等)上的准确性显著优于 MiniCPM-V、DeepSeek-VL、Yi-VL 等开源大模型; 在专用教育评测数据集上, VELM 的准确性相较于基础模型 Qwen2-VL 提升了 9.78%。

## 关键词

大语言模型; 多模态; 智慧教育; RAG 技术

中图分类号: TP39

文献标志码: A

doi:10.11959/j.issn.2096-0271.2025056

## Optimization and application of vision-based large models in educational scenarios

XU Yuepeng<sup>1,2</sup>, XU Chaidi<sup>3</sup>, GUO Jinjun<sup>1</sup>, JIANG Yunqiao<sup>4</sup>, WANG Shijia<sup>1</sup>, LIU Yao<sup>1,2,5</sup>

1. Data Science and Engineering, East China Normal University, Shanghai 200062, China
2. Laboratory for Advanced Computing and Intelligence Engineering, Wuxi 214000, China
3. Shanghai Kongdi Technology Co., Ltd., Shanghai 201204, China
4. Moore Threads Intelligent Technology (Chengdu) Co., Ltd., Chengdu 610031, China
5. MoE Engineering Research Center of Software/Hardware Co-Design Technology and Application (East China Normal University), Shanghai 200062, China

## Abstract

With the rapid advancement of artificial intelligence technology, LLMs have achieved significant success across various

fields. However, their application in the field of education domain still faces challenges such as difficulties in processing multimodal data, insufficient response accuracy, and limited information delivery methods. To address these issues, a VELM was proposed. VELM was trained on multimodal public educational datasets and specialized educational datasets, and combined with model optimization techniques, VELM not only enhances response quality in educational scenarios but also optimizes and reduces computational resource consumption. Additionally, RAG technology was utilized to ensure accuracy and richness in generated content. In terms of deployment and application, VELM was implemented through the Dify platform, enabling flexible multi-end deployment, including WeChat mini programs, Web cloud platforms, and localized deployment, meeting the diverse needs of different educational scenarios. Evaluation experiments demonstrated that VELM significantly outperformed open-source large models such as MiniCPM-V, DeepSeek-VL, and Yi-VL on standard benchmark datasets like Mathvista, OCRBench, and MMMU. On specialized educational evaluation datasets, the accuracy of VELM was improved by 9.78% compared to the base model Qwen2-VL.

**Key words**  
large language model, multimodal, smart education, RAG technology

## 0 引言

随着人工智能技术的不断发展，大语言模型（large language model, LLM）逐渐涌现并重塑着人们的工作和生活。然而，其在教育领域的应用仍面临诸多挑战。首先，教育数据通常是多模态的，包括文本、图像、音频、视频等，而传统大语言模型往往无法有效处理这些多模态数据，难以提供全面的学习支持。其次，传统大语言模型通常需要人工设计提示或指令（Prompt），难以适应教育场景中多样化、自动化的任务需求，例如复杂推理题的分析等，这在一定程度上限制了其准确性和实用性。

为了解决上述问题，本文提出了一种视觉教育大模型（visual education large model, VELM）。VELM的一个重要优势是其对多模态数据的处理能力较强。相比于只能处理文本数据的传统大模型，VELM能够同时处理文本、图像等多模态

输入，提供更全面的学习支持。VELM还通过对大规模教育数据集的微调与优化，提高了模型在教育场景中的适应性与性能。利用量化、剪枝等模型优化技术，VELM在保证高效推理的同时，大幅降低了计算资源的消耗，使其能够在资源有限的教育环境中得到更广泛的应用。与此同时，VELM还结合检索增强生成（retrieval-augmented generation, RAG）技术生成更适合教育场景的提示词。通过这种方式，VELM可以更精准地处理逻辑推理、分析和解答问题，解决传统大模型在教育任务中的“幻觉”问题，并且提高模型对复杂任务的适应性。

本文主要贡献如下：

- （1）多模态教育数据集构建：通过整合开源数据集、原始教材和自制教育数据集，构建了一个适用于多模态教育场景的高质量数据集，为视觉教育大模型的优化与应用提供了基础。
- （2）基于多模态数据集的模型微调与优化：使用多模态教育数据集对模型进行微调，并且结合量化、剪枝、FlashAttention等技术，有效降低了模型计算开销，提高了

模型的性能，确保模型在离线环境下也可高效运行。

(3) 基于 RAG 技术的智能化部署：RAG 技术帮助模型实现了高效的信息检索，增强了其在实际教学任务中的智能化应用能力，并且实现了在微信小程序、Web 云端和本地化部署等多种场景的应用。

## 1 相关工作

### 1.1 视觉大模型

近年来，视觉大模型（vision large model, VLM）在人工智能领域取得了飞速发展，越来越多的科技公司和研究机构推出了具有代表性的视觉大模型，进一步推动了视觉任务的性能提升和应用拓展。其中，OpenAI 推出的多模态模型 GPT-4<sup>[1]</sup>，是一款具备强大视觉和语言处理能力的模型。它不仅能够理解和生成自然语言，还支持对图像的分析 and 描述，在视觉问答、图像生成和图文结合任务等方面展现了卓越的性能。此外，阿里巴巴达摩院推出的 Qwen-VL 模型<sup>[2]</sup>将视觉与语言的能力进行了有机融合。Qwen-VL 能够处理复杂的图文任务，如图像中的物体检测、描述生成以及图文问答等，并在大规模多模态数据的训练中展现了卓越的视觉理解与生成能力<sup>[3]</sup>。清华大学和面壁智能联合发布的 MiniCPM-V 模型<sup>[4]</sup>则代表了视觉大模型轻量化的探索方向。作为一款面向视觉任务的轻量级多模态预训练模型，MiniCPM-V 在保持较小参数规模的前提下，依然能够在视觉与跨模态任务中取得良好的性能。它通过创新的预训练策略和模型结构设计，实现了对视觉与语言信息的有效融合，具备处理图像分类、目标检测、图文匹配等

多种任务的能力。

### 1.2 PEFT 微调技术

参数高效微调（parameter-efficient fine-tuning, PEFT）是一种针对大规模预训练模型的先进微调策略。其主要目标在于通过调整极少量模型参数，实现优异的性能表现，同时显著削减训练成本及存储需求。PEFT 的核心策略聚焦于对模型的一小部分参数进行精细化调整，而非全盘更新，从而有效降低训练过程中的计算负担及存储空间需求<sup>[5]</sup>。PEFT 技术的主流实现路径如下。

(1) Adapter：此策略在预训练模型的每一层级中嵌入小巧的适配器模块。训练阶段，仅针对这些适配器模块的参数进行优化，而维持原模型参数的稳定性。此举大幅缩减了待更新参数的数量，同时借助适配器模块实现了对特定任务的灵活适配<sup>[6]</sup>。

(2) LoRA (low-rank adaptation)：该方法将模型中的部分权重矩阵分解为两个低秩矩阵，进而在微调阶段仅针对这些低秩矩阵进行调整，避免了对原始模型完整权重的全面修改。该方法在确保模型性能的同时，显著减少了待更新参数的规模<sup>[7]</sup>。

(3) Prefix Tuning：该方法在输入序列的前端附加一段可训练的前缀，并在训练过程中仅针对该前缀的参数进行优化，而保持预训练模型其余部分的参数固定不变。通过精细调整前缀内容，模型能够迅速适应新任务的需求<sup>[8]</sup>。

(4) Prompt Tuning：Prompt Tuning 与 Prefix Tuning 类似，但其仅在输入中添加少量可训练的提示词，并在训练阶段专注于这些提示词参数的优化，而

不涉及模型其他部分的调整<sup>[9]</sup>。

1.3 RAG 技术

RAG<sup>[10]</sup>是一种优化大型语言模型输出的方法，使其能够在生成响应之前引用训练数据之外的权威知识库。LLM 使用海量数据进行训练，拥有数十亿个参数，能够执行诸如回答问题、翻译语言和完成句子等任务。RAG 在 LLM 强大功能的基础上，访问特定领域或组织的内部知识库，而不用重新训练模型，进一步提升了其输出的相关性、准确性和实用性。这是一种经济高效的改进方法，适用于各种情境。RAG 包含 3 个主要过程：检索、增强和生成。

(1) 检索

根据用户的查询内容，从外部知识库获取相关信息。具体而言，将用户的查询嵌入模型，并转换为向量，以便与向量数据库中存储的相关知识进行比对。通过相似性搜索，找出与查询最匹配的前  $K$  个数据。

(2) 增强

将用户的查询内容和检索到的相关知识一起嵌入一个预设的提示词模板中。

(3) 生成

将经过检索增强的提示词内容输入大型语言模型，以生成所需的输出。

1.4 Dify 平台

随着 LLM 的快速发展，如何高效地将这些模型应用于实际场景成了一项巨大挑战。尽管 LLM 展现出了强大的生成能力，但将其部署到生产环境中仍然存在诸多挑战，包括复杂的系统集成、资源管理和持续维护等<sup>[11]</sup>。为了应对这些问题，研究者开发了 Dify 平台。Dify 平台融合了后端即

服务（backend as a service, BaaS）和大语言模型操作（large language model operation, LLMOp）的理念，旨在提供一个一站式解决方案，使开发者能够快速搭建生产级的生成式 AI 应用。Dify 支持多种大型语言模型，如 Llama、QWen、Baichuan、ChatGPT 等，并与多个模型供应商合作，确保开发者可以根据具体需求选择最适合的模型。此外，Dify 还提供了一系列工具和服务，包括模型部署、性能监控、自动化更新等，从而显著降低了开发和维护的成本与复杂性。Dify 的出现不仅有助于加速生成式 AI 应用的落地，也为未来的 LLM 开发提供了解决方案。

2 数据集收集与构建

在教育场景中，数据集的质量和丰富性是训练大模型的关键，特别是在处理涉及多模态信息（如文本、图像和视频）和教育领域特定任务的模型中，数据集的作用尤为重要。数据集来源主要分为 3 个部分：开源的教育数据集、原始教材资料，以及自制数据集。其中，自制数据集是本研究的创新点之一，通过收集大量教育资料构建数据集，进一步增强了模型的适应性。自制数据集流程如图 1 所示。

2.1 开源的教育数据集

本研究使用了来自多个开源平台的数据集，这些数据集涵盖了教育领域中的常见任务，如阅读理解、知识问答等。这些数据集也包含了来自多个学科的纯文本数据集和多模态数据集。例如，来自 OpenData 的 Eedi 数据集提供了包含两个学年（2018 年 9 月—2020 年 5 月）学生对

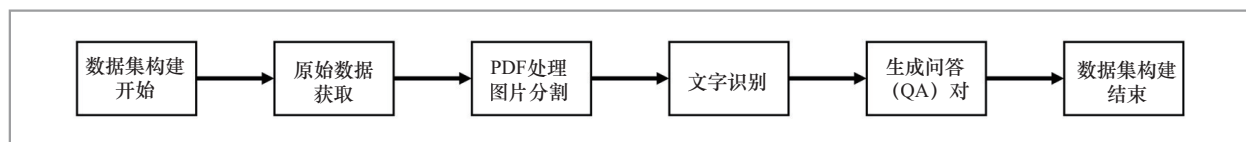


图1 自制数据集流程

Eedi 数学问题的回答，主要用于提高模型的数学能力。ChemData700K 汇集了9项化学核心任务，730K个高质量问答的化学能力指令微调数据集。此外，OCR-VQA数据集提供了超过20万张图像和相关的问答（QA）对，主要用于提高模型的视觉能力<sup>[12]</sup>。

这些开源数据集为视觉大模型提供了广泛的基础训练数据，使其能够具备较强的图文理解与处理能力。然而，开源数据集的局限在于内容的通用性，它们并未完全针对特定的教育场景设计，因此在实际应用中，特别是面向不同年级和学科的教育任务时，还需要通过其他手段进一步增强模型的能力。这也是本研究中结合原始教材和自制数据集的必要性所在。

## 2.2 原始教材数据

为了进一步提高模型回答的正确率，本研究使用了大量的原始教材数据作为RAG技术的知识库来源。教材不仅包含丰富的课程知识，还涵盖了多学科、多年级的教学内容，这使得它在教育领域的问答和内容生成任务中具有强大的支持作用。这些数据集包含从教材中提取的文本内容，涵盖了各个学科的知识点、定义、定理和例题等。本研究从教材中提取了总计超过50 000页的文本数据。

原始教材数据的来源如下。（1）与华东师范大学出版社合作，获取了经过授权的教材电子版，这些教材涵盖了从小学到

高中的各个年级和学科。（2）从2个在线教育资源平台（国家中小学智慧教育平台、华东师范大学大夏学堂）收集了公开可用的教材数据。这2个平台由教育机构或政府部门运营，确保了数据的质量和合法性。（3）本研究还使用笔者团队自主编写的教材，字数为100.1万字，目前教材使用人数已超过3万人。

RAG技术将视觉大模型与检索系统结合起来，使模型不仅依赖于内部参数生成内容，还可以动态地从外部知识库中获取相关信息，为模型提供准确且全面的知识支持。具体而言，模型在回答学生的学习问题时，能够先从教材中检索出相关的章节、段落或知识点，然后基于检索到的信息生成准确的回答。这种方法尤其适用于处理复杂的教育任务，例如，解释科学概念、推导数学公式或提供历史事件的背景信息。这不仅提高了模型回答的准确性，也保证了生成内容的可信度。

## 2.3 自制数据集

自制数据集是本研究的核心贡献之一。为了解决现有开源数据集在某些教育任务中数据局限的问题，本研究构建了一个专门的教育数据集。这个数据集的构建流程涉及数据获取、预处理、自动化问答生成等多个步骤，最终形成了一个高质量的问答（QA）对数据集，用于训练和评估视觉大模型。训练数据集任务列表见表1。

表1 训练数据集任务列表

| 训练数据集   | 视觉问答(VQA)任务 | 数量/条    |
|---------|-------------|---------|
| 教育训练数据集 | —           | 约50 000 |
|         | 单选题         | 约35 000 |
|         | 多选题         | 约10 000 |
|         | 简答题         | 约5 000  |

数据集构建流程如下。

(1) 数据获取

本研究首先爬取了大量公开的试卷数据。这些试卷通常以PDF格式提供，包含大量的选择题、判断题和简答题，覆盖了

多个学科的考试内容。共收集了约2 000份试卷。平台上的数据确保了自制数据集的真实性。

(2) PDF文件处理和图像分割

爬取到的试卷通常是扫描的PDF文档。为了便于进一步处理，首先将PDF文件转变为图片进行保存，之后采用目标检测方法自动检测和分割试卷中的题目、题号、选项等关键信息，将其从图片中框取出来。共处理了超过6 000张图像。图片框取结果示例如图2所示。

(3) 文字识别

针对分割后的试卷题目信息，本研究使用了通用OCR理论（general OCR

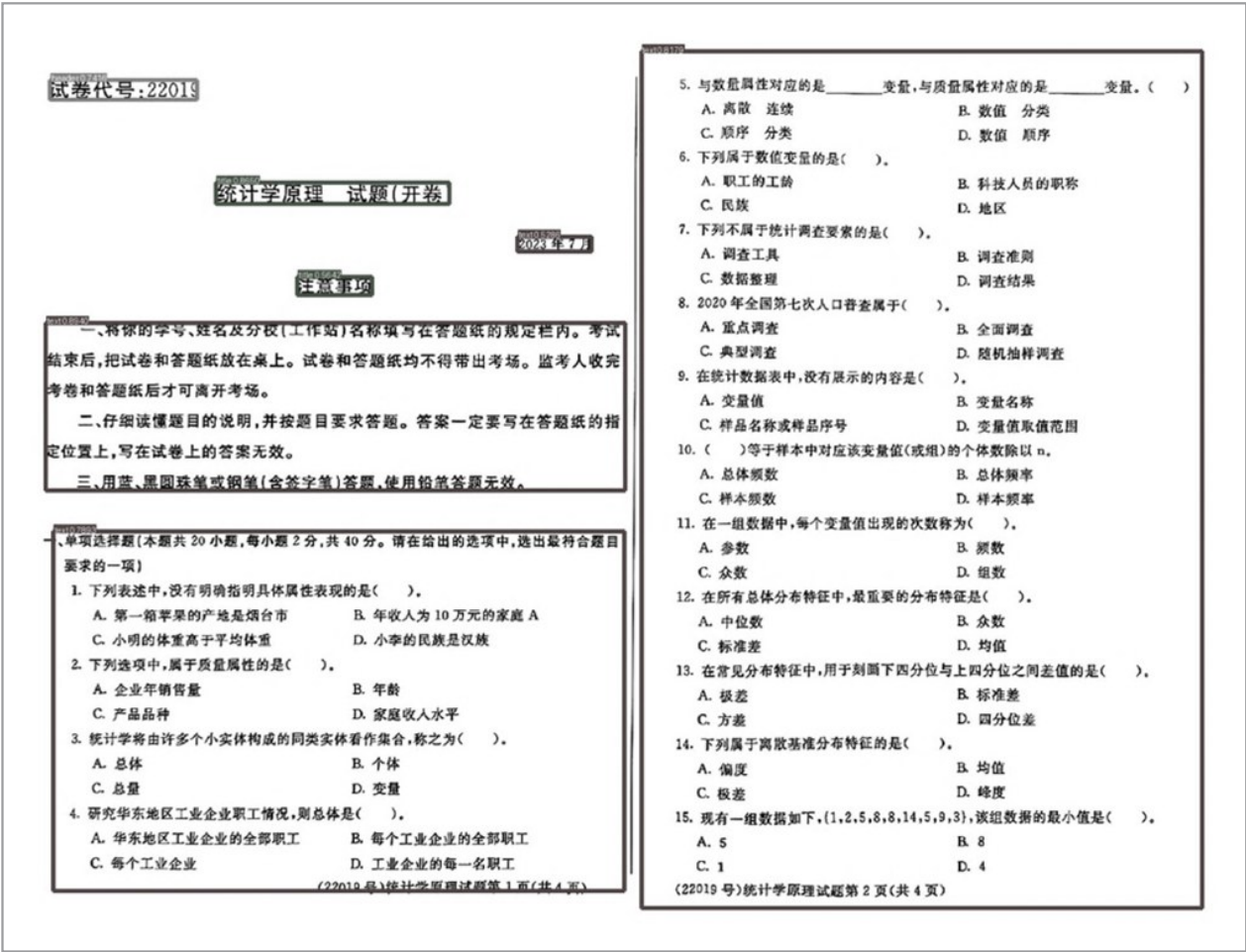


图2 图片框取结果示例

theory, GOT) 技术进行文字识别, 将试卷中的题目和文本信息转换为可编辑的 TXT 格式, 转换后的文本数据主要以“题号: 题目”的形式进行存储。共成功识别并转换了约 50 000 道题目。

(4) 生成 QA 对数据集

通过调用 ChatGPT-4o 模型, 本研究为每个题目生成了对应的答案, 并将这些题目和答案组合成标准化的 QA 对。该过程不仅满足了数据集的格式, 并且提高了数据集的质量。最终生成的 QA 对数据集包含约 50 000 个高质量的 QA 对。

应多样化的教育需求。为了进一步提高模型微调的效率, 还采用了量化和 FlashAttention 等优化技术, 对模型进行压缩和加速, 使其在资源受限的环境中仍能高效运行。在推理阶段, 结合 RAG 技术, 模型在回答时能够实时从外部知识库中检索相关的背景信息, 确保生成的答案不仅准确, 而且具有逻辑性和可靠的知识支持。最终通过 Dify 平台部署, 使 VELM 可以灵活地适应各种特定的教育任务场景, 如个性化辅导、智能题目生成以及多模态问答等任务。此外, 还提供了 3 种应用方式, 分别是小程序端、Web 云端和本地化部署方式。

### 3 视觉教育大模型设计

#### 3.1 视觉教育大模型架构设计

本研究基于 Qwen-VL-7B 设计了 VELM, 以应对教育场景中的多模态数据处理和复杂推理任务。VELM 整体架构如图 3 所示。首先, 通过 3 种不同的教育数据集对模型进行了微调, 使其能够适

#### 3.2 算法流程概述

本节详细介绍模型的算法流程, 包括输入输出的设计, 以及如何通过 RAG 技术将输入数据与 Query 结合生成增强 Prompt, 从而提升模型的推理性能和输出质量。

模型的输入由多模态数据构成, 涵盖

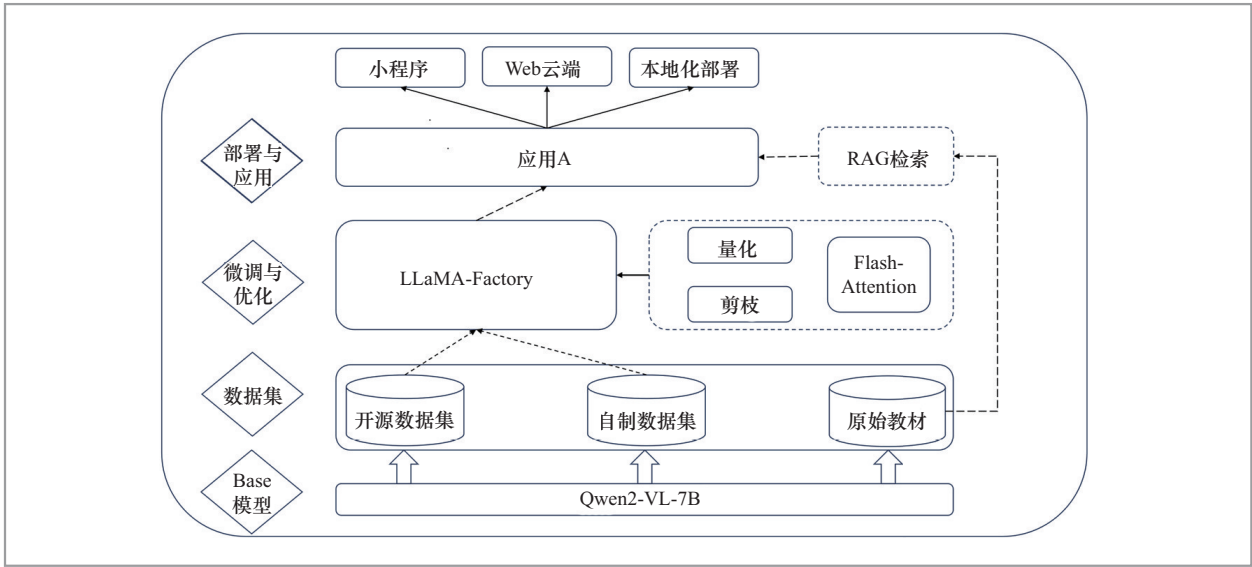


图3 VELM整体架构

文本和图像两种类型。文本输入通常包括用户提出的问题描述或学习材料，例如，任务指令、参考文献等内容，用于明确模型需要解决的目标。对于多模态场景，图像输入作为文本的补充，可以是教育领域的图表、流程图或其他相关的图片信息。模型的输出则是针对输入问题的详细答案，能够根据用户需求提供逻辑清晰、内容准确的回答，尤其在整合多模态信息的场景下表现尤为突出。

在处理输入数据的过程中，充分利用了检索增强生成技术，将用户的 Query 与外部资料相结合以生成增强版 Prompt。具体而言，模型首先通过向量化技术将用户的 Query 转换为向量表示，并在向量数据库中匹配检索最相关的资料。检索到的资料通常较多，因此需要进一步进行 Rerank 操作，即根据与 Query 的相关性和资料的质量对检索结果重新排序，以确保最优质、最相关的资料能够被优先选取。

经过筛选的资料随后与预设的 Prompt 模板结合，形成具有明确上下文的提示输入。Prompt 的设计是 RAG 技术的关键，它能够有效将 Query 与检索到的资料整

合，确保模型对任务语境的深度理解。本研究采用的模板如下：“请根据以下相关资料来清晰准确地回答问题，假如与你的知识有所冲突，以相关资料的内容为准。相关资料：{result}Query：{query}”。通过这种结合方式，生成的增强 Query 更具语义完整性和上下文关联性。增强后的 Query 最终被输入预训练语言模型（如 VELM）中进行推理，从而生成高质量的答案输出。RAG 技术流程示意图如图 4 所示。

3.3 模型微调与优化

本研究采用 LLama-Factory 框架对 Base 模型 Qwen2-VL-7B 进行了微调。LLama-Factory 是一款高效的训练和微调工具，适用于处理大规模语言模型和多模态任务<sup>[13]</sup>。本研究使用前文所提到的两类数据集对其进行微调：一部分是开源的多模态教育数据集，涵盖了文本、图像及问答数据；另一部分是本文的自制教育数据集，针对小众的学科进行生成，使模型获得更广泛的知识。数据集

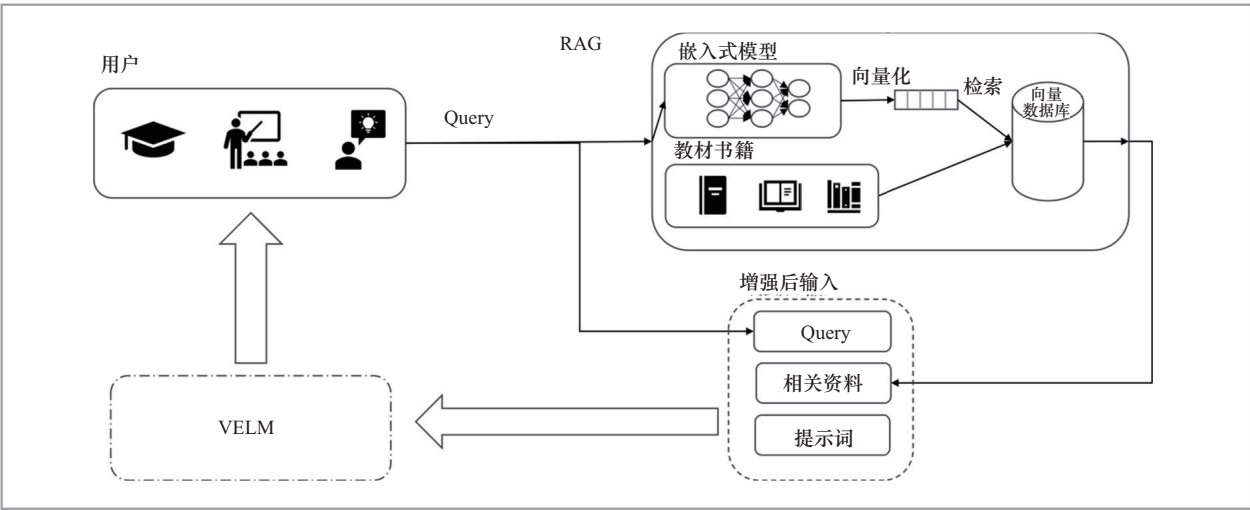


图4 RAG技术流程示意图

的多样性确保模型能够在多个教育任务中实现高效的推理与生成，QLoRA 训练过程如算法 1 所示。

算法 1 Large Model Fine-tuning with QLoRA

---

```

1: Input:
2:  $M$ : Qwen2VL
3:  $D$ : Training dataset (Query-Answer pairs  $(x, y)$ )
4:  $L$ : Loss function
5:  $\eta$ : Learning rate
6:  $E$ : Number of epochs
7:  $r$ : LoRA rank (low-rank dimension)
8:  $\lambda$ : Scaling factor for LoRA updates
9:  $q$ : Quantization level (bit-width for quantization)
10: Output:
11:  $M_{\text{fine-tuned}}$ : VELM
12: Step 1: Initialize QLoRA parameters
13: for each weight matrix  $W \in M$  do
14:   Freeze  $W$  in the pre-trained model.
15:   Introduce two low-rank trainable matrices:
16:    $A \in R^{d \times r}$  (down-projection)
17:    $B \in R^{r \times d}$  (up-projection)
18:   Quantize the weight matrices  $A$  and  $B$  to  $q$ -bits:
19:    $A_q \leftarrow \text{Quantize}(A, q)$ 
20:    $B_q \leftarrow \text{Quantize}(B, q)$ 
21:   Compute the LoRA-adapted weight as:
22:    $\hat{W} = W + \lambda A_q B_q$ 

```

---



---

```

23: end for
24: Step 2: Train QLoRA parameters
25: for each epoch  $e \in \{1, 2, \dots, E\}$  do
26:   for each batch  $(x, y) \in D$  do
27:     Forward pass through the model using  $W$ .
28:     Compute the loss:  $L = L(M(x), y)$ 
29:     Backpropagate the gradients with respect to  $A_q$  and  $B_q$ .
30:     Update  $A_q$  and  $B_q$  using gradient descent:
31:      $A_q \leftarrow A_q - \eta \cdot \frac{\partial L}{\partial A_q}$ 
32:      $B_q \leftarrow B_q - \eta \cdot \frac{\partial L}{\partial B_q}$ 
33:   end for
34: end for
35: Step 3: Fine-tuned model output
36: Return the fine-tuned model  $M_{\text{fine-tuned}}$  with updated QLoRA parameters.

```

---

本研究的微调方法选择 QLoRA 方法<sup>[14]</sup>。该方法通过低秩适配矩阵优化参数调节，减少了训练所需的资源量，并提升了模型的调优效率。为了进一步提升训练的效率 and 减少资源消耗，本研究采用了混合精度训练 (FP16) 技术，这不仅减少了显存占用，还加快了模型的训练速度。

在推理优化方面，本研究采用了 Flash-Attention 技术<sup>[15]</sup>。该技术优化了 Transformer 架构中的自注意力机制，大幅提升了训练和推理阶段的计算效率，减少了内存访问瓶颈，确保了在处理大规模输入时的高效计算。通过 Flash-Attention 技术，VELM 在面对大量复杂的

教育推理任务时表现出了更快的响应速度。

在模型优化方面，本研究使用了量化和剪枝技术。首先，通过量化将浮点运算从 FP32 降为 FP16 甚至 INT8，这显著减少了模型的存储和计算开销，同时尽量保证模型的精度不受显著影响。此外，剪枝技术通过识别并移除模型中不必要的神经元和连接，将模型的参数数量压缩到更小的规模。经过这些优化，VELM 在离线环境中也能保持高效运行，特别适用于资源有限的教育场景，如远程教育设备或断网情况下的部署。

3.4 部署与应用

为了确保 VELM 在各种实际教育场景中的高效应用，本研究将 VELM 部署到了 Dify 平台。Dify 平台是一个集成 BaaS 和 LLMOps 的工具，能够简化大语言模型的部署、管理和持续优化流程。借助 Dify 平台的强大支持，不仅实现了 VELM 在云端和本地的无缝部署，还大大简化了 VELM 的管理、监控与更新，为用户提供稳定、高效的使用体验。

当用户使用 VELM 进行学习辅导或问题解答时，会通过 RAG 技术提升生成答案的准确性和信息丰富度。具体流程是：当用户输入问题后，系统会从预先构建的外部教育知识库中进行检索，找到与用户问题相关的数据。这些检索到的内容会被转换为参考信息，并与用户输入一起作为输入数据输送到 VELM 模型中进行推理。通过这种方式，VELM 不仅依赖于自身的内部知识，还能够实时获取并利用最新的外部信息，大幅提升答案的准确性和知识的全面性。

此外，本研究还针对不同的使用场景，特别设计了灵活多样的部署与访问方式。VELM 可以通过以下3种方式为用户提供服务。

(1) 小程序

VELM 在小程序端的部署主要面向移动端用户，特别是学生群体。学生可以随时随地通过手机或平板电脑，轻松访问 VELM，进行问题提问或是拍照问答，获得个性化的即时学习指导。小程序的设计简洁直观，用户只需输入问题或图片，VELM 就会通过 RAG 进行检索，并提供精准的答案。该方案操作灵活且便捷，适用于碎片化学习和日常作业辅导，极大提升了学习的便利性和互动性。

(2) Web 云端

VELM 的 Web 云端部署主要面向教师、教育工作者以及研究人员。用户可以通过浏览器访问 VELM，进行批量的教育资源生成、复杂问题解答以及知识点解析等任务。Web 云端的界面更加专业化，支持更高复杂度的任务处理，如智能试题生成、多模态教材分析等。教师可以利用这一平台快速生成针对性练习题，或通过 VELM 进行多模态数据（如图文、视频）的辅助教学，帮助学生更好地理解复杂的学习内容。

(3) 本地化部署

除了直接的用户端应用，VELM 还可以通过本地化部署来使用。用户可以在下载 VELM 模型后将其部署到本地，将 VELM 的智能推理能力集成到其他教育平台、应用或学习管理系统中。这种灵活的部署方式，允许教育科技公司或学校定制自己的智能辅导系统，或基于 VELM 的强大功能开发创新型教育产品。

4 实验结果与分析

4.1 实验设置

本研究实验在拥有两张 NVIDIA

GeForce RTX 4090 显卡的服务器上进行。针对实验中使用的 Qwen2-VL-7B 模型，基于 LLama-Factory 框架进行了微调优化。在超参数的选择上，经过多次实验调优，最终确定批大小（batch size）为 64，以确保在计算资源允许的情况下，能够有效提升训练速度。同时，将学习率（learning rate）设置为  $5 \times 10^{-4}$ ，这一设置在初步实验中表现出良好的收敛性，能够有效平衡模型的训练速度与精度。此外，训练周期（epoch）设置为 5 轮次，综合考虑了训练时间与模型性能之间的平衡，以避免过拟合现象的发生。数据集包括前文所说的自定义数据集和开源的教育数据集。为了确保数据质量和模型的泛化能力，所有数据在输入模型之前均进行了标准化处理。实验设置见表 2。

4.2 模型评测

本次模型评测包括在专用教育评测数据集与标准评测数据集上的表现。

(1) 专用教育评测数据集

该数据集是为了评估模型的学科内容理解和多模态识别能力而设计的。其中共包含生物、化学、历史、科学、数学等 7 个不同学科，包含了多种任务，不同任务的评价指标也有所不同。专用教育评测数据集学科题目占比如图 5 所示。

(2) 标准评测数据集

在模型性能评估中，标准评测数据集为模型的基本效果提供了重要的参考。这些数据集通常由业界广泛认可，涵盖多种任务并提供统一的评估指标。常见的标准评测数据集包括 MMMU、OCRBench 和 MathVista。MMMU 用于评估多模态语言理解能力，涵盖阅读理解、文本生成等多种任务，为模型提供多角度的评测<sup>[16]</sup>。

表 2 实验设置

| 类别            | 设置                      |
|---------------|-------------------------|
| 操作系统          | Ubuntu 20.04            |
| 内存            | 16 GB × 2               |
| 显卡            | NVIDIA GeForce RTX 4090 |
| 模型            | Qwen2-VL-7B             |
| batch size    | 64                      |
| learning rate | $5 \times 10^{-4}$      |
| epoch         | 5                       |
| 数据集           | 自定义数据集、开源的教育数据集         |

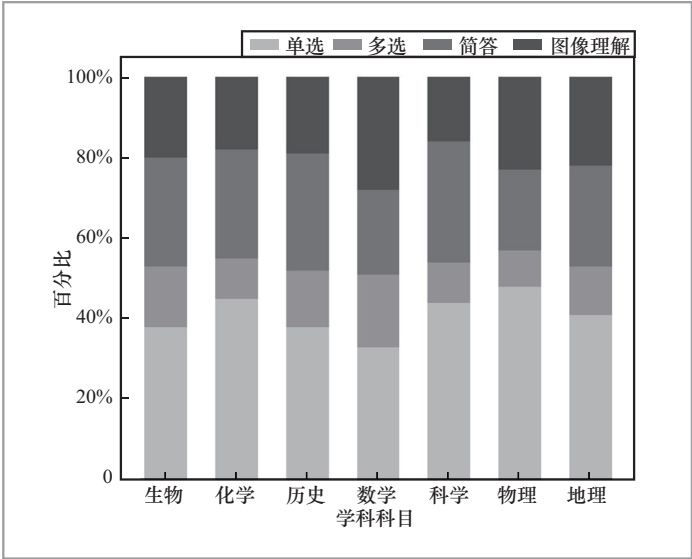


图 5 专用教育评测数据集学科科目占比

OCRBench 针对大规模多模态大模型 (large multimodal models, LMMs) 的全面光学字符识别 (optical character recognition, OCR) 评估基准，旨在评估模型在处理各种文本相关视觉任务中的能力<sup>[17]</sup>。它包含 29 个数据集，是目前最全面的 OCR 评估基准之一。MathVista 用于评估大型语言模型和大规模多模态模型在视觉上下文中的数学推理能力<sup>[18]</sup>。这个基准测试解决了评估这些模型在结合数学推

理和视觉理解任务中表现的不足。

专用教育评测数据集任务列表与标准评测数据集部分任务列表见表3。

4.3 实验方案

为了全面评估模型的性能，在专用教育评测数据集和标准评测数据集上进行了对比实验。自定义数据集包含约30 000条样本，任务类型涵盖单选题、多选题、简答题等；标准评测数据集使用了公开的任

务数据集，确保模型在通用任务中的表现得到充分验证。

对比实验中的模型包括VELM、Qwen2-VL-7B（Base模型）、Yi-VL-6B、Deepspeek-VL-7B以及MiniCPM-Llama3-V 2.5模型。Qwen2-VL-7B是一个适用于图像描述生成、图像问答、对图像内容进行推理和解释等任务的视觉模型<sup>[19]</sup>。Yi-VL6B是零一万物Yi系列模型家族发布的多模态大模型，拥有较强的视觉理解能力<sup>[20]</sup>。DeepSpeek-VL是幻方量化

| 表3 评测数据集任务列表         |                     |             |          |
|----------------------|---------------------|-------------|----------|
| 评测数据集                | 任务                  | 数量          | 评测指标     |
| 专用教育评测数据集            | —                   | 29 000条     | —        |
|                      | 单选题                 | 12 000条     | Accuracy |
|                      | 多选题                 | 3 000条      | Accuracy |
|                      | 简答题                 | 8 000条      | Bleu-4   |
|                      | 图像理解                | 6 000条      | Bleu-4   |
| 标准评测数据集<br>MMMU      | —                   | 超过11 500种任务 | —        |
|                      | STEM(科学、技术、工程和数学)任务 | —           | Accuracy |
|                      | 人文学科任务              | —           | Accuracy |
|                      | 社会科学任务              | —           | Accuracy |
|                      | 其他任务                | —           | Accuracy |
| 标准评测数据集<br>MathVista | —                   | 共29个数据集     | —        |
|                      | VQA任务               | —           | Bleu-4   |
|                      | 数学单词问题              | —           | Accuracy |
|                      | 几何问题解决              | —           | Accuracy |
|                      | 逻辑推理任务              | —           | Bleu-4   |
| 标准评测数据集<br>OCR Bench | 科学推理任务              | —           | Bleu-4   |
|                      | —                   | —           | —        |
|                      | 文本识别                | —           | Bleu-4   |
|                      | 场景文本视觉问答            | —           | Accuracy |
|                      | 文档导向视觉问答            | —           | Accuracy |
|                      | 关键信息提取              | —           | Bleu-4   |
|                      | 手写数学表达式识别           | —           | Bleu-4   |

旗下深度求索公司研发的模型系列，对代码、数学问题的理解能力较强。MiniCPM-Llama3.5V 是面壁智能公司与清华大学共同开发的面向图文理解的端侧多模态大模型系列。该系列模型接受图像和文本输入，并提供高质量的文本输出，它们分别代表了多模态任务和中文语言生成任务的不同基线。实验重点评估生成文本的相似性、回答的准确度以及图片理解等多模态处理能力。实验分为两轮：第一轮在自定义数据集上进行，测试模型在复杂教育场景任务中的表现；第二轮在标准评测数据集上进行，评估其在通用多模态任务中的适应性与泛化能力。最后汇总各项指标，全面评估模型的实际应用效果。

## 4.4 实验结果

### (1) 专用教育评测数据集

在专用教育评测数据集任务中，VELM 模型在多方面表现优异，相比于 Base 模型得分提升了 9.78%。它在理解和回答特定学科内容方面展现出了显著的优势，能够提供高质量的回答，充分证明了其在教育场景中的高适应性。此外，在选择题的处理上，它具备较强的准确性，能够选出正确答案，显著降低了漏选和误选的可能。相较于其他模型，VELM 在处理图像理解任务时也具备了极高的准确性，能够有效提取和理解图像中的文本信息，表现出良好的多模态理解能力。

### (2) 开源的教育数据集

在 OCR-Bench 数据集和涉及复杂数学推理的 MathVista 任务中，VELM 的表现均略逊于最优模型，其相较于 Base 模型的表现仅下降了 2% 和 3%。尽管这一差距不大，但反映出 VELM 在处理这些特定任务时的表现略显不足。一个可能的原因

是，VELM 的参数规模相对较小，其能够学习到的知识容量有限，这使得当模型针对某一能力（如教育场景下的回答能力）进行优化时，往往会导致其他能力不足（如跨场景的文字识别和多步数学推理），从而引发性能的轻微下降。这种能力间的此消彼长现象源于模型规模限制对知识存储与泛化能力的约束。尽管如此，VELM 的整体表现仍然接近最优，未来还需针对模型规模与知识分配策略之间的关联进行优化，以提升其在复杂任务中的表现。

在 MMMU 任务中，VELM 的整体表现相较于 Base 模型有所提升，尤其是在多学科背景的任务中展现出更高的稳定性。具体而言，VELM 不仅能够适应不同领域的知识需求，还能够灵活应对多样化的回答场景。例如，在跨学科的推理任务中，VELM 可以准确捕捉到问题的上下文和细节，提供清晰且准确的回答。此外，VELM 的优势还体现在其对多类型问题的兼容性上，针对文本理解与生成任务、图像识别与理解任务、多模态问答任务等，它都能提供稳定的输出。总体来说，VELM 作为通用教育大模型，在不同任务中的表现显示了它具备较强的跨领域适应能力和推理深度，特别是在需要多模态信息整合和复杂逻辑推理的场景下。模型实验结果见表 4。其中加粗的为最佳结果，加下划线的为次佳结果。在不同评测集上的得分对比如图 6 所示。

## 4.5 部署与展示

### (1) 网页端

为了方便地展示和应用 VELM 模型，本研究通过 Dify 平台将其部署在网页端。Dify 平台提供了高效的模型部署工具，能够轻松集成和管理多模态模型。利用

表4 模型实验结果

| 模型          | Size/B | 教育评测数据<br>集得分 | 标准评测数据集得分  |                               |                     |
|-------------|--------|---------------|------------|-------------------------------|---------------------|
|             |        |               | OCRBench   | MathVista <sub>testmini</sub> | MMMU <sub>val</sub> |
| Qwen2-VL    | 7      | 63.3          | <b>866</b> | <b>58.2</b>                   | <u>54.1</u>         |
| MiniCPM-V   | 8.5    | 60.8          | 725        | 54.3                          | 45.8                |
| Yi-VL       | 6.7    | 51.7          | 290        | 28.8                          | 40.3                |
| DeepSeek-VL | 7.3    | 49.7          | 435        | 36.8                          | 38.3                |
| VELM        | 7      | <b>69.5</b>   | <u>848</u> | <u>56.4</u>                   | <b>56.2</b>         |

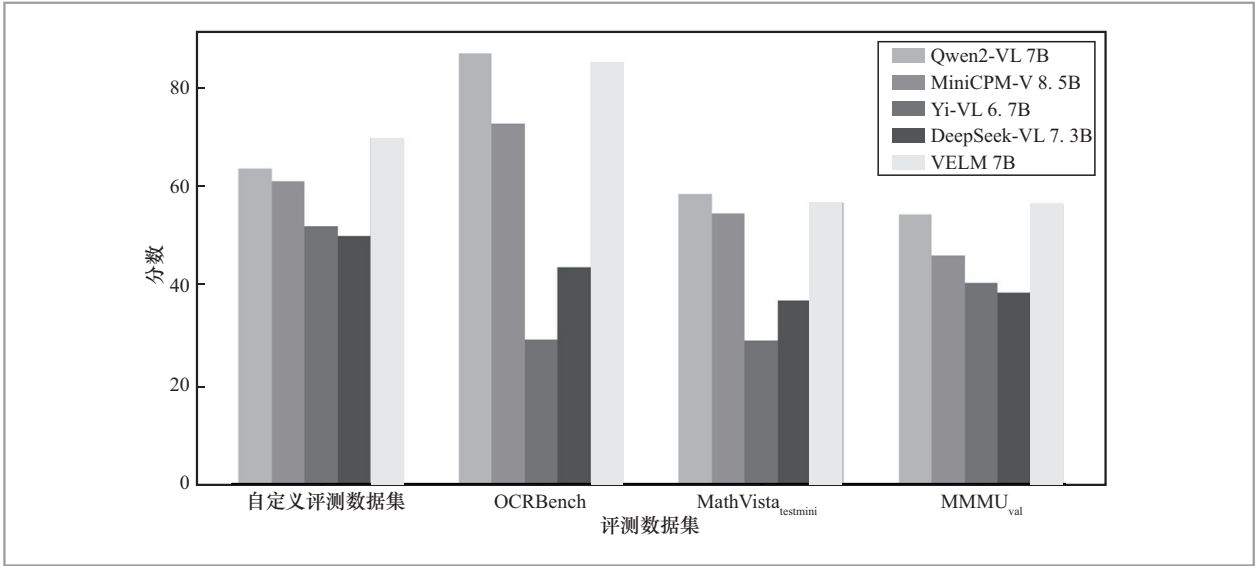


图6 模型得分对比

Gradio框架构建交互式界面，用户可以在浏览器中进行多模态输入（如文本、图像），并实时查看模型的输出结果。网页端展示界面简洁直观，允许用户上传自定义知识库，并且可以根据知识库的内容自动生成QA对。QA对自动生成展示如图7所示。与此同时，在网页端还可以实现文本与图像在线问答。图像生成问答展示如图8所示。

（2）小程序端

为了拓展模型的实际应用场景，本研究设计并部署了“水杉码园智能助教”小

程序，用户可以通过微信等平台直接访问这一智能助手，实现多模态交互，如文本问答和图像识别等功能。

具体而言，小程序提供了简洁易用的界面，帮助用户快速访问不同功能模块。在首页，界面顶部展示了用户欢迎信息，并设有一个搜索框，方便用户输入问题以获取答案。同时，小程序分为多个功能区域，例如“水杉助教”和“论文助手”模块，分别对应日常问题解答和学术支持的需求。小程序还提供了“魔法咒语”部分，列出了常见问题的快捷入口，用户可以快



图7 QA对自动生成展示



图8 图像生成问答展示

速找到“如何有效记笔记”“大学生职业规划”等常见问题的解决方案。

此外，界面下方还设有“灵感创意”板块，内含诸如“考试复习”“知心朋友”之类的推荐功能，帮助用户高效地制订学习计划或者在学习过程中获得激励和支持。整个小程序的设计目标在于提升用户的交互体验，让用户能够随时随地通过手机获得高效的智能问答服务。小程序端的首页及问答页面展示如图9所示。

5 结束语

本研究通过设计和实现 VELM，为智能教育领域中的多模态数据处理、复杂推理任务和个性化教学提供了新的解决方案。在模型的设计过程中，本文以 Qwen2-VL-7B 为基础模型，结合了 RAG 技术，确保模型在面对教育领域中多样化、复杂



图9 小程序端的首页及问答页面展示

化的任务时，能够给出准确的解答。此外，借助 LLama-Factory 平台的高效训练框架，本文对模型进行了微调，利用量化、剪枝等优化技术显著提升了模型的推理效率和部署灵活性，确保 VELM 能够适应在线和离线的多种应用场景。

在模型的部署和应用方面，VELM 通过 Dify 平台进行了灵活的多端部署，支持小程序、Web 云端和本地化部署 3 种形式，满足了不同用户群体的需求。无论是面向学生的个性化学习辅导，还是为教师提供智能试题生成与教学资源分析，VELM 都展现了强大的实际应用价值。

尽管本研究取得了一些成果，但也存在一定的局限性。例如，在特定的教育任务中，模型的性能可能受到模型规模大小、任务复杂性和多模态数据噪声的限制。此外，由于算力的限制，本文无法在更大的模型规模上来进行实验。未来的工作将集中于进一步扩展 VELM 的功能，优化其在实际场景中的应用效果，并结合更多的教育数据集来提升模型的通用性和准确性。本文也将对 VELM 在不同教育场景中的实际表现进行深入分析，以便为后续的研究和应用提供更全面的数据支持。

## 参考文献：

- [1] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report[EB]. arXiv preprint, 2023, arXiv:2303.08774.
- [2] WU C F, YIN S M, QI W Z, et al. Visual ChatGPT: talking, drawing and editing with visual foundation models[J]. arXiv preprint, 2023: 2303.04671.
- [3] BAI J Z, BAI S, YANG S S, et al. Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond[EB]. arXiv preprint, 2023, arXiv: 2308.12966.
- [4] YAO Y, YU T Y, ZHANG A, et al. MiniCPM-V: a GPT-4V level MLLM on your phone[EB]. arXiv preprint, 2024, arXiv: 2408.01800.
- [5] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21(1): 5485-5551.
- [6] HOULSBY N, GIURGIU A, JASTRZEB-SKI S, et al. Parameter-efficient transfer learning for NLP[C]//Proceedings of the 36th International Conference on Machine Learning. PMLR, 2019: 2790-2799.
- [7] HU E J, SHEN Y L, WALLIS P, et al. Lora: low-rank adaptation of large language models[EB]. arXiv preprint, 2021, arXiv: 2106.09685.
- [8] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation[EB]. arXiv preprint, 2021, arXiv: 2101.00190.
- [9] LESTER B, AL-RFOU R, CONSTANT N, et al. The power of scale for Parameter-efficient prompt tuning[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021: 3045-3059.
- [10] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 9459-9474.
- [11] PRANAV R, JEREMY I, ZHU K, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning[EB]. arXiv preprint, 2017, arXiv: 1711.05225.
- [12] WEI H R, LIU C L, CHEN J Y, et al.

General OCR theory: towards OCR-2.0 via a unified end-to-end model[EB]. arXiv preprint, 2024, arXiv: 2409.01704.

[13] ZHENG Y W, ZHANG R C, ZHANG J H, et al. LlamaFactory: unified efficient fine-tuning of 100+ language models[EB]. arXiv preprint, 2024, arXiv: 2403.13372.

[14] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. Qlora: Efficient finetuning of quantized llms[C]// Proceedings of the 37th International Conference on Neural Information Processing System. New York: Curran Associates Inc., 2023: 10088 - 10115.

[15] DAO T, FU D, ERMON S, et al. Flashattention: fast and memory-efficient exact attention with IO-Awareness[EB]. arXiv preprint, 2022, arXiv: 2205.14135.

[16] YUE X, NI Y S, ZHENG T Y, et al. MMMU: a massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 9556-9567.

[17] LIU Y L, LI Z, HUANG M X, et al. OCRBench: on the hidden mystery of OCR in large multimodal models[J]. Science China, 2024, 67(12): 220102

[18] LU P, BANSAL H, XIA T, et al. Math-Vista: evaluating mathematical reasoning of foundation models in visual contexts[EB]. arXiv preprint, 2023, arXiv: 2310.02255.

[19] WANG P, BAI S, TAN S N, et al. Qwen2-VL: enhancing vision-language model's perception of the world at any resolution[EB]. arXiv preprint, 2024, arXiv: 2409.12191.

[20] YOUNG A, Chen B, LI C, et al. Yi: open foundation models by 01. AI[EB]. arXiv preprint, 2024, arXiv: 2403.04652.

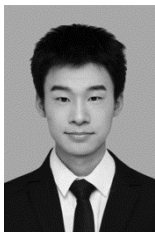
作者简介



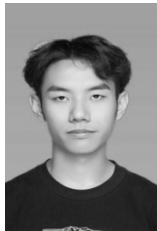
许跃蓬（2001-），男，华东师范大学数据科学与工程学院硕士生，主要研究方向为智能教育、大模型性能优化。



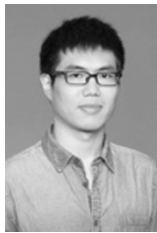
徐柴迪（1997-），男，现就职于上海孔楝科技有限公司，主要研究方向为文本生成、高性能计算。



郭晋军（2002-），男，华东师范大学数据科学与工程学院硕士生，主要研究方向为智能教育、大模型性能优化。



姜云桥（1996-），男，现就职于摩尔线程智能科技有限公司，主要研究方向为并行计算、高性能计算。



王仕嘉（1987-），男，华东师范大学数据科学与工程学院工程师，主要研究方向为智能教育等。



刘垚（1981-），男，博士，华东师范大学数据科学与工程学院副教授，主要研究方向为并行计算、智能计算等。

收稿日期: 2024-11-05  
通信作者: 刘垚, liuyao@cc.ecnu.edu.cn  
基金项目: 国家自然科学基金项目(No.42375146);国家重大科技基础设施项目(No.2024-EL-PT-000737);先进计算与智能工程国家级重点实验室基金项目(No.2023-LYJJ-01-006);光合基金项目(No.202407013820)  
Foundation Items: The National Natural Science Foundation of China (No.42375146), The National Key Scientific and Technological Infrastructure Project (No.2024-EL-PT-000737), The Fund of Laboratory for Advanced Computing and Intelligence Engineering (No. 2023-LYJJ-01-006), GHfund A (No.202407013820)